Receiver-Operating-Characteristic Analysis Reveals Superiority of Scale-Dependent Wavelet and Spectral Measures for Assessing Cardiac Dysfunction

Stefan Thurner,[1] * Markus C. Feurstein,[1] † Steven B. Lowen,[1] and Malvin C. Teich[1,2]
[1] *Department of Electrical and Computer Engineering, Boston University, Boston, Massachusetts 02215, USA*
[2] *Departments of Biomedical Engineering and Physics, Boston University, Boston, Massachusetts 02215, USA*

Receiver-operating-characteristic (ROC) analysis was used to assess the suitability of various heart rate variability (HRV) measures for correctly classifying electrocardiogram records of varying lengths as normal or revealing the presence of heart failure. Scale-dependent HRV measures were found to be substantially superior to scale-independent measures (scaling exponents) for discriminating the two classes of data over a broad range of record lengths. The wavelet-coefficient standard deviation at a scale near 32 heartbeat intervals, and its spectral counterpart near 1/32 cycles per interval, provide reliable results using record lengths just minutes long. A jittered integrate-and-fire model built around a fractal Gaussian-noise kernel provides a realistic, though not perfect, simulation of heartbeat sequences.
PACS number(s) 87.10.+e, 87.80.+s, 87.90.+y

Though the notion of using heart rate variability (HRV) analysis to assess the condition of the cardiovascular system stretches back some 40 years, its use as a noninvasive clinical tool has only recently come to the fore [1]. A whole host of measures, both scale-dependent and scale-independent, have been added to the HRV armamentarium over the years.

One of the more venerable among the many scale-dependent measures in the literature is the interbeat-interval (R-R) standard deviation $\sigma_{\rm int}$ [2]. The canonical example of a scale-independent measure is the scaling exponent α_S of the interbeat-interval power spectrum, associated with the decreasing power-law form of the spectrum at sufficiently low frequencies f: $S(f) \propto f^{-\alpha_S}$ [1,3]. Other scale-independent measures have been developed by us [4–6], and by others [7–9]. One of the principal goals of this Letter is to establish the relative merits of these two classes of measures, scale-dependent and scale-independent, for assessing cardiac dysfunction.

One factor that can confound the reliability of a measure is the nonstationarity of the R-R time series. Multiresolution wavelet analysis provides an ideal means of decomposing a signal into its components at different scales [10–12], and at the same time has the salutary effect of eliminating nonstationarities [13,14]. It is therefore ideal for examining both scale-dependent and scale-independent measures; it is in this latter capacity that it provides an estimate of the wavelet scaling exponent α_W [6].

We recently carried out a study [6] in which wavelets were used to analyze the R-R interval sequence from a standard electrocardiogram (ECG) database [15]. Using the wavelet-coefficient standard deviation $\sigma_{\rm wav}(m)$, where $m = 2^r$ is the scale and r is the scale index, we discovered a critical scale window near $m = 32$ interbeat intervals over which it was possible to perfectly discriminate heart-failure patients from normal subjects. The presence of this scale window was confirmed in an Israeli-Danish study of diabetic patients who had not yet developed clinical signs of cardiovascular disease [16]. These two studies [6,16], in conjunction with our earlier investigations which revealed a similar critical scale window in the *counting* statistics of the heartbeat [4,5,17] (as opposed to the time-*interval* statistics considered here), lead to the recognition that scales in the vicinity of $m = 32$ enjoy a special status. This conclusion has been borne out for a broad range of analyzing wavelets, from Daubechies 2-tap (Haar) to Daubechies 20-tap [10,18] (higher order analyzing wavelets are suitable for removing polynomial nonstationarities [11]). It is clear that scale-dependent measures [such as $\sigma_{\rm wav}(32)$] substantially outperform scale-independent ones (such as α_S and α_W) in their ability to discriminate patients with certain cardiac dysfunctions from normal subjects (see also [18,19]).

The reduction in the value of the wavelet-coefficient standard deviation $\sigma_{\rm wav}(32)$ that leads to the scale window occurs not only for heart-failure patients [6], but also for heart-failure patients with atrial fibrillation [18], diabetic patients [16], heart-transplant patients [16,19], and in records preceeding sudden cardiac death [6,19]. The depression

*now at Institut für Kernphysik, TU-Wien, Wiedner Hauptstrasse 8-10, A-1040 Vienna, Austria
†now at Dept. of Industrial Information Processing, WU-Wien, Pappenheimgasse 35/3/5, A-1200 Vienna, Austria

of $\sigma_{\rm wav}(32)$ at these scales is likely associated with the impairment of autonomic nervous system function. Baroreflex modulations of the sympathetic or parasympathetic tone typically lie in the range 0.04–0.09 Hz (11–25 sec), which corresponds to the time range where $\sigma_{\rm wav}(m)$ is reduced.

The perfect separation achieved in our initial study of 20-h Holter-monitor recordings endorses the choice of $\sigma_{\rm wav}(32)$ as a useful diagnostic measure. The results of most studies are seldom so clear-cut, however. When there is incomplete separation between two classes of subjects, as observed for other less discriminating measures using these identical long data sets [7,8], or when our measure is applied to large collections of out-of-sample or reduced-length data sets [19], an objective means for determining the relative diagnostic abilities of different measures is required.

ROC Analysis.– Receiver-operating-characteristic (ROC) analysis [20] is an objective and highly effective technique for assessing the performance of a measure when it is used in binary hypothesis testing. This format provides that a data sample be assigned to one of two hypotheses or classes (e.g., normal or pathologic) depending on the value of some measured statistic relative to a threshold value. The efficacy of a measure is then judged on the basis of its sensitivity (the proportion of pathologic patients correctly identified) and its specificity (the proportion of control subjects correctly identified). The ROC curve is a graphical presentation of sensitivity versus 1−specificity as a threshold parameter is swept (see Fig. 1).

The area under the ROC curve serves as a well-established index of diagnostic accuracy [20]; a value of 0.5 arises from assignment to a class by pure chance whereas the maximum value of 1.0 corresponds to perfect assignment (unity sensitivity for all values of specificity). ROC analysis can be used to choose the best of a host of different candidate diagnostic measures by comparing their ROC areas, or to establish for a single measure the tradeoff between reduced data length and misidentifications (misses and false positives) by examining ROC area as a function of record length (see Fig. 2). A minimum record length can then be specified to achieve acceptable classification. Because ROC analysis relies on no implicit assumptions about the statistical nature of the data set [18,20], it is more reliable and appropriate for analyzing non-Gaussian time series than are measures of statistical significance such as p-value and d' which are expressly designed for signals with Gaussian statistics [18]. Moreover, ROC curves are insensitive to the units employed (e.g., spectral magnitude, magnitude squared, or log magnitude); ROC curves for a measure M are identical to those for any monotonic transformation thereof such as M^x or $\log(M)$. In contrast the values of d', and its closely related cousins, change under such transformations. Unfortunately, this is not always recognized which leads some authors to specious conclusions [21].

Scale-Dependent vs Scale-Independent Measures.– Wavelet analysis provides a ready comparison for scale-dependent and scale-independent measures since it reveals both. ROC curves constructed using 75,821 R-R intervals from each of the 24 data sets (12 heart failure, 12 normal) [15], are presented in Fig. 1 (left) for the wavelet measure $\sigma_{\rm wav}(32)$ (using the Haar wavelet) as well as for the wavelet measure α_W. It is clear from Fig. 1 that the area under the $\sigma_{\rm wav}(32)$ ROC curve is unity, indicating perfect discriminability. This scale-dependent measure clearly outperforms the scale-independent measure α_W which has significantly smaller area. These results are found to be essentially independent of the analyzing wavelet [6].

We now use ROC analysis to quantitatively compare the tradeoff between reduced record length and misidentifications for this standard set of heart-failure patients using three scale-dependent and three scale-independent measures. In the first category are the wavelet-coefficient standard deviation $\sigma_{\rm wav}(32)$, its spectral counterpart $S(1/32)$ [18,22], and the interbeat-interval standard deviation $\sigma_{\rm int}$. In the second category, we consider the wavelet scaling exponent α_W, the spectral scaling exponent α_S, and a scaling exponent α_D calculated according to detrended fluctuation analysis (DFA) [8].

In Fig. 2 (left) we present ROC area, as a function of R-R interval record length, using these six measures. The area under the ROC curves forms the rightmost point in the ROC area curves. The file sizes are then divided into smaller segments of length L. The area under the ROC curve is computed for the first such segment for all 6 measures, and then for the second segment, and so on for all segments of length L. From the $L_{\rm max}/L$ values of the ROC area, the mean and standard deviation are computed. The lengths L employed range from $L = 2^6 = 64$ to $L = 2^{16} = 65,536$ in powers of two.

The best performance is achieved by $\sigma_{\rm wav}(32)$ and $S(1/32)$, both of which attain unity area (perfect separation) for sufficiently long R-R sequences. Even for fewer than 100 heartbeat intervals, corresponding to *just a few minutes of data*, these measures provide excellent results (in spite of the fact that both diurnal and nocturnal records are included). $\sigma_{\rm int}$ does not perform quite as well. The worst performance, however, is provided by the three scaling exponents α_W, α_S, and α_D, confirming our previous findings [4–6]. Moreover, results obtained from the different power-law estimators differ widely [23], suggesting that there is little merit in the concept of a single exponent, no less a "universal" one [21], for characterizing the human heartbeat sequence. In a recent paper Amaral et al. [21] conclude exactly the opposite, that the scaling exponents provide the best performance. This is because they improperly make use of the Gaussian-based measures d^2 and η, which are closely related to d', rather than ROC analysis. These same

authors [21] also purport to glean information from higher moments of the wavelet coefficients, but such information is not reliable because estimator variance increases with moment order. The results presented here accord with those obtained in a detailed study of 16 different measures of HRV [18]. There are vast differences in the time required to compute these measures however: for 75,821 interbeat intervals, $\sigma_{\rm wav}(32)$ requires the shortest time (20 msec) whereas DFA(32) requires the longest time (650,090 msec).

It will be highly useful to evaluate the relative performance of these measures for other records, both normal and pathologic. In particular the correlation of ROC area with severity of cardiac dysfunction should be examined.

An issue of importance is whether the R-R sequences, and therefore the ROC curves, arise from deterministic chaos [9]. We have carried out a phase-space analysis in which *differences* between adjacent R-R intervals are embedded. This minimizes correlation in the time series which can interfere with the detection of deterministic dynamics. The results indicate that the behavior of the underlying R-R sequences, both normal and pathological, appear to have stochastic rather than deterministic origins [18].

Generating a realistic heartbeat sequence.– The generation of a mathematical point process that faithfully emulates the human heartbeat could be of importance in a number of venues, including pacemaker excitation. Integrate-and-fire (IF) models, which are physiologically plausible, have been developed for use in cardiology. Berger et al. [24], for example, constructed an integrate-and-fire model in which an underlying rate function was integrated until it reached a fixed threshold, whereupon a point event was triggered and the integrator reset. Improved agreement with experiment was obtained by modeling the stochastic component of the rate function as band-limited fractal Gaussian noise (FGN), which introduces scaling behavior into the heart rate, and setting the threshold equal to unity [5]. This fractal-Gaussian-noise integrate-and-fire (FGNIF) model has been quite successful in fitting a whole host of interval- and count-based measures of the heartbeat sequence for both heart-failure patients and normal subjects [5]. However, it is not able to accommodate the differences observed in the behavior of $\sigma_{\rm wav}(m)$ for the two classes of data.

To remedy this defect, we have constructed a jittered version of this model which we dub the fractal-Gaussian-noise jittered integrate-and-fire (FGNJIF) model [23]. The occurrence time of each point of the FGNIF is jittered by a Gaussian distribution of standard deviation J. Increasing the jitter parameter imparts additional randomness to the R-R time series at small scales, thereby increasing $\sigma_{\rm wav}$ at small values of m and, concomitantly, the power spectral density at large values of the frequency f. The FGNJIF simulation does a rather good job of mimicing patient and control data for a number of key measures used in heart-rate-variability analysis. The model is least successful in fitting the interbeat-interval histogram $p_\tau(\tau)$, particularly for heart-failure patients. This indicates that that a mechanism other than jitter for increasing $\sigma_{\rm wav}$ at low scales should be sought [18].

It is of interest to examine the global performance of the FGNJIF model using the collection of 24 data sets. To achieve this we carried out FGNJIF simulations using parameters comparable with the actual data and constructed simulated ROC curves for the measures $\sigma_{\rm wav}(32)$ and α_W as shown in Fig. 1 (right). Similar simulations for ROC area versus record length are displayed in Fig. 2 (right) for the six experimental measures considered. Overall, the global simulations (right-hand side of Fig. 1 and 2) follow the trends of the data (left-hand side of Fig. 1 and 2) reasonably well, with the exception of $\sigma_{\rm int}$. This failure is linked to the inability of the simulated results to emulate the observed interbeat-interval histograms. It will be of interest to consider modifications of the FGNIF model that might bring the simulated ROC curves into better accord with the data-based curves.

---

[1] M. Malik, J. T. Bigger, A. J. Camm, R. E. Kleiger, A. Malliani, A. J. Moss, and P. J. Schwartz, *Euro. Heart J.* **17**, 354–381 (1996) [*Circulation* **93**, 1043–1065 (1996)].

[2] M. M. Wolf, G. A. Varigos, D. Hunt, and J. G. Sloman, *Med. J. Australia* **2**, 52–53 (1978).

[3] M. Kobayashi and T. Musha, *IEEE Trans. Biomed. Eng.* **BME-29**, 456–457 (1982).

[4] R. G. Turcott and M. C. Teich, *Proc. SPIE* **2036**, 22–39 (1993).

[5] R. G. Turcott and M. C. Teich, *Ann. Biomed. Eng.* **24**, 269–293 (1996).

[6] S. Thurner, M. C. Feurstein, and M. C. Teich, *Phys. Rev. Lett.* **80**, 1544–1547 (1998).

[7] C.-K. Peng, J. Mietus, J. M. Hausdorff, S. Havlin, H. E. Stanley, and A. L. Goldberger, *Phys. Rev. Lett.* **70**, 1343–1346 (1993).

[8] C.-K. Peng, S. Havlin, H. E. Stanley, and A. L. Goldberger, *Chaos* **5**, 82–87 (1995).

[9] J. B. Bassingthwaighte, L. S. Liebovitch, and B. J. West, *Fractal Physiology* (Oxford Univ. Press, New York, 1994).

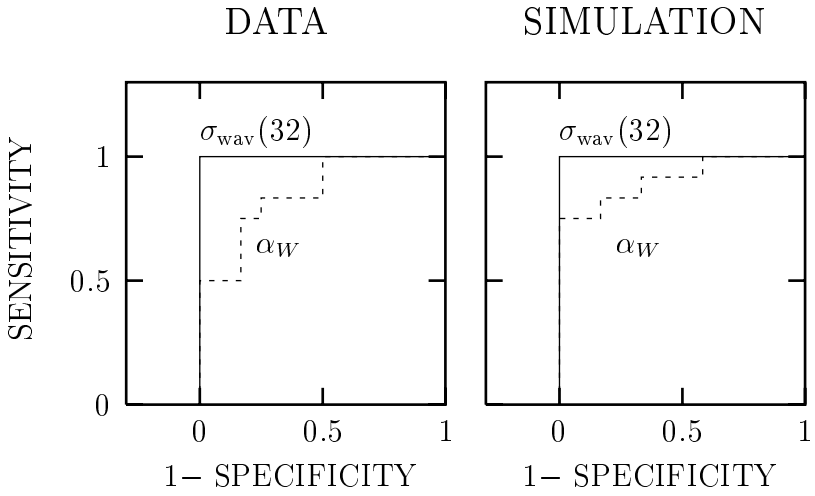
[10] I. Daubechies, *Ten Lectures on Wavelets* (Society for Industrial and Applied Mathematics, Philadelphia, PA, 1992).

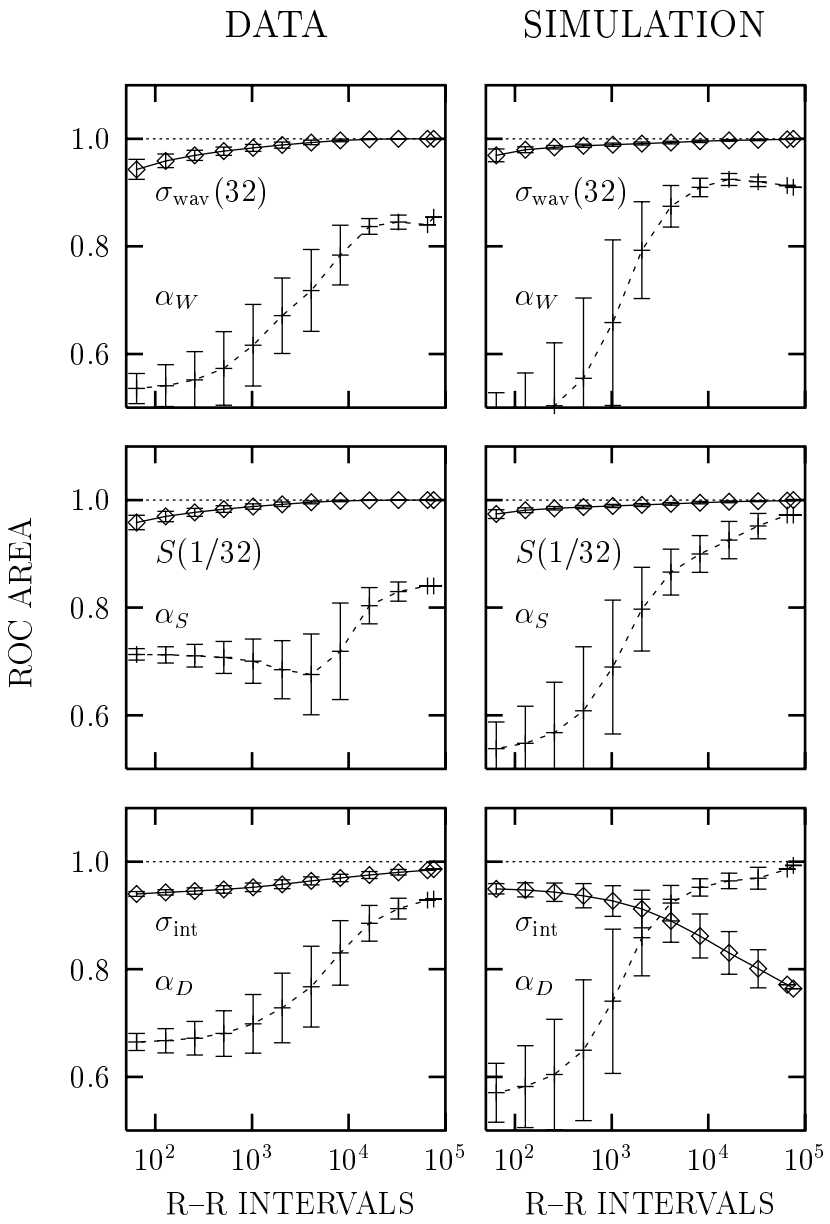
[11] *Wavelets in Medicine and Biology*, edited by A. Aldroubi and M. Unser (CRC Press, Boca Raton, FL, 1996).

[12] *Time Frequency and Wavelets in Biomedical Signal Processing*, edited by M. Akay (IEEE Press, Piscataway, NJ, 1997).
[13] A. Arneodo, G. Grasseau, and M. Holschneider, *Phys. Rev. Lett.* **61**, 2281-2284 (1988).
[14] M. C. Teich, C. Heneghan, S. B. Lowen, and R. G. Turcott, in *Wavelets in Medicine and Biology* (CRC Press, Boca Raton, FL, 1996), pp. 383–412.
[15] The R-R recordings used in Ref. [6] were drawn from the Beth-Israel Hospital (Boston, MA) Heart-Failure Database comprising 12 records from normal subjects (age 29–64 years, mean 44 years) and 15 records from severe congestive heart-failure patients (age 22–71 years, mean 56 years). The recordings were made with a Holter monitor digitized at a fixed value of 250 Hz. Three of the 15 heart-failure patients suffered from atrial fibrillation. The analysis reported in this Letter eliminates these three atrial-fibrillation patients, and makes use of the first 75,821 sinus-rhythm interbeat intervals from the remaining 12 normal and 12 congestive heart-failure records. A detailed characterization of each of the records is presented in Table 1 of Ref. [5]. The conclusions reported in this Letter remain intact even when the three atrial-fibrillation patients are included, as shown elsewhere [18].
[16] Y. Ashkenazy, M. Lewkowicz, J. Levitan, H. Moelgaard, P. E. Bloch Thomsen, and K. Saermark, *Fractals* **6**, 197–203 (1998).
[17] M. C. Teich, *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.* **18**, 1128–1129 (1996).
[18] M. C. Teich, S. B. Lowen, B. M. Jost, K. Vibe-Rheymer, and C. Heneghan, in *Nonlinear Biomedical Signal Processing* (IEEE Press, Piscataway, NJ, 1999), in press.
[19] M. C. Teich, *Proc. Int. Conf. IEEE Eng. Med. Biol. Soc.* **20** (#3), 1136–1141 (1998).
[20] J. A. Swets, *Science* **240**, 1285–1293 (1988).
[21] L. A. Nunes Amaral, A. L. Goldberger, P. Ch. Ivanov, and H. E. Stanley, *Phys. Rev. Lett.* **81**, 2388–2391 (1998).
[22] C. Heneghan, S. B. Lowen, and M. C. Teich, "Analysis of spectral and wavelet-based measures used to assess cardiac pathology," submitted to International Conference on Acoustic, Signal, and Speech Processing (ICASSP) (1999).
[23] S. Thurner, S. B. Lowen, M. C. Feurstein, C. Heneghan, H. G. Feichtinger, and M. C. Teich, *Fractals* **5**, 565–595 (1997).
[24] R. D. Berger, S. Askelrod, D. Gordon, and R. J. Cohen, *IEEE Trans. Biomed. Eng.* **BME-33**, 900–904 (1986).

FIG. 1. ROC curves (sensitivity vs 1−specificity) for two wavelet-based measures: $\sigma_{\rm wav}(32)$ which is scale-dependent and α_W which is scale-independent. *Left*: ROC curves obtained using all 24 data records, each comprising 75,821 interbeat intervals [15]. The scale-dependent measure outperforms the scale-independent one since its ROC area is greater. *Right*: Comparable result obtained using simulations for the fractal-Gaussian-noise jittered integrate-and-fire (FGNJIF) model.

FIG. 2. Diagnostic accuracy (area under ROC curve) *vs* data length (number of R-R intervals) for three scale-dependent and three scale-independent measures (mean ± one standard deviation). An area of unity corresponds to the correct assignment of each patient to the appropriate class. *Left*: $\sigma_{\rm wav}(32)$ and $S(1/32)$ provide excellent performance, attaining unity area (perfect separation) for 32,768 (or more) R-R intervals. These measures continue to perform well even as the number of R-R intervals decreases below 100, corresponding to record lengths just minutes long. The performance of $\sigma_{\rm int}$ is seen to be slightly inferior. In contrast, all three scale-independent measures perform poorly. *Right*: Similar results are obtained using 24 simulations of the FGNJIF model, with the exception of $\sigma_{\rm int}$ (see text).


DATA
SIMULATION
SENSITIVITY
1
0.5
0
$\sigma_{\mathrm{wav}}(32)$
α_W
0
0.5
1
1− SPECIFICITY
$\sigma_{\mathrm{wav}}(32)$
α_W
0
0.5
1
1− SPECIFICITY


DATA
SIMULATION
ROC AREA
1.0
0.8
0.6
$\sigma_{\rm wav}(32)$
α_W
$S(1/32)$
α_S
$\sigma_{\rm int}$
α_D
10^2
10^3
10^4
10^5
R–R INTERVALS
R–R INTERVALS